An Evaluation of OFCCP's Equal Opportunity Survey

Prepared for:
Office of Federal Contract Compliance Programs,
Employment Standards Administration,
U.S. Department of Labor

Prepared by:
Abt Associates Inc.
55 Wheeler Street
Cambridge, MA 02138

Prepared under:
Contract No. GS10F0086K
Purchase Order No. B9635233

Table of Contents

Chapter 1. Introduction

In 1999 and 2000 the Office of Federal Contract Compliance Programs (OFCCP), in the Employment Standards Administration of the U.S. Department of Labor, developed the Equal Opportunity Survey (EO Survey). The aim was "to meet three major program objectives: (1) to increase compliance with equal opportunity requirements by improving contractor self-awareness and encouraging self-evaluations; (2) to improve the deployment of federal government resources toward contractors most likely to be out of compliance; and, (3) to increase agency efficiency by building on the tiered-review process already authorized by OFCCP's regulatory reform efforts, thereby allowing better resource allocation" (OFCCP 2000).

After cognitive testing and field testing, OFCCP selected a sample of nearly 7,000 contractor establishments to test the utility of the EO Survey and mailed it out to them in April 2000. Part of this sample (nearly 3,000 establishments, the "3K sample") had been identified as potential subjects of evaluation but not recently evaluated, and the remainder (the "4K sample") had completed an evaluation within the prior two fiscal years. Unfortunately, much of the evaluation data had little bearing on the forms of noncompliance that involve systemic discrimination.

Equally unfortunate, but perhaps to be expected in light of the evaluation data, the survey data were subject to influence by the evaluations. A study of the predictive power of the EO Survey by Bendick and Egan Economic Consultants, Inc. noted the problematic relation between the 4K sample and the evaluation data and concluded that "The EO Survey data collected in the April 2000 wave does not offer circumstances in which the full predictive power of the survey can be revealed." (Bendick et al. 2000, p. 20). Bendick et al. explained that the timing of data collection posed a particular handicap because the effects of corrective actions arising from the compliance evaluations might have seriously distorted the statistical estimates of predictive relationships.

In a separate report based on the 3K sample, OFCCP concluded "that the EO Survey data and the analytic model are able to distinguish between those establishments with a high potential for findings of violations or deficiency and those with low probability for findings of violations or deficiencies" (OFCCP 2000). The focus of the OFCCP report, however, was on a wide variety violations or deficiencies that included relatively minor paperwork deficiencies, rather than on systemic discrimination. Moreover, the OFCCP study did not account for the high number of false positives that would result from that preliminary model.

The first full implementation took place in FY2001, when the EO Survey was sent to nearly 50,000 establishments (one half of the contractor universe from the file of EEO-1 forms submitted for FY1999), sampled at a rate of 50% in strata based on size (but excluding establishments in the sample for 2000). The data from the responses were entered into a database. No establishments, however, were selected for compliance evaluations on the basis of those data.

In 2002 OFCCP contracted with Abt Associates Inc. to design and draw a sample of approximately 10,000 establishments from a subset of the establishments that had EEO-1 contractor records from FY2000. That sample (described in detail in Chapter 2) formed the basis

for the 2002 EO Survey, which was mailed in December 2002. Abt Associates also designed and drew a subsample (also described in Chapter 2) of establishments that would undergo compliance evaluations linked to their responses on the EO Survey. Together, the data from the 2002 EO Survey and the results of those compliance evaluations were intended to provide a basis for a study of targeting, for compliance evaluations, establishments that are likely to be involved in systemic discrimination. Abt Associates also developed a plan for such a study of targeting.

This report describes the study carried out by Abt Associates, under a subsequent contract with the Department of Labor. As mentioned above, Chapter 2 discusses the design of the sample for the 2002 EO Survey and the subsample for compliance evaluations. Chapter 3 reviews the processes of data collection and gives the results. Chapter 4 describes the development of potential predictor variables, derived from the data collected in the 2002 EO Survey. Chapter 5 brings together those predictor variables and the findings from the compliance evaluations to build a logistic regression model for targeting systemic discrimination. It also evaluates the model and its predictive ability. Finally, Chapter 6 summarizes the study and states its main conclusions.

Addendum

OFCCP's review of the draft of this report led to requests for additional tabulations, analyses, and discussion. Appendix E contains the memorandum submitted by Abt Associates in response to those requests. Where appropriate, the final version of this report refers to material in that memorandum, which contains Tables A through E.

Chapter 2. Sample Design

The data for the study came from responses to the 2002 EO Survey and from reviews of a subsample of the establishments selected for the 2002 EO Survey. This chapter describes the initial frame of establishments and the procedures for selecting the sample and the subsample.

Sampling Frame

The target population consisted of a subset of the 95,961 establishments with EEO-1 contractor records for FY2000. The subset excluded the following categories:

- Establishments that were sent EO Surveys the previous year.
- Establishments that the OFCCP reviewed within the last two years (FY2001 and FY2002).
- Establishments associated with a parent company for which the OFCCP has approved a Functional Affirmative Action Program.
- Any establishment that had the same parent company as an establishment that had asserted that the OFCCP lacked jurisdiction (for reasons that comprised five categories).
- A small number of establishments that had very questionable records.
- Establishments that were among the 6,863 to which EO Surveys were sent in April 2000, in connection with the pilot study.
- All establishments of two large companies that have traditionally contested jurisdiction and were not sent EO Surveys on the previous round.

The resulting subset contained 26,451 establishments. A sample of approximately 10,000 establishments was drawn from this sampling frame, according to an allocation among a detailed set of strata.

Strata and Allocation

In preparing for this study, it was important to select the sample in a way that ensured coverage of the sampling frame as a whole and of specific subsets of the sampling frame that had relevant combinations of characteristics. Among the variables available in the EEO-1 contractor records for FY2000, three characteristics seemed desirable to cover:

- Region, derived by taking the first letter of the 2-letter code for the district office (6 values: B, C, D, E, F, and I)

- Industry, defined by 12 groupings of SIC code (shown in Table 2.1)

- Size, given by the variable SIZETYPE (4 values: < 150 employees, 150 to 299, 300 to 499, and 500 or more).

The possible combinations of these three characteristics defined a total of 288 (= 6 x 12 x 4) strata. The sampling frame contained a nonzero number of establishments in each of these 288 strata. The table in Appendix A gives the frequency distribution of the 26,451 establishments over the 288 strata.

Table 2.1 Industry Groupings, Based on 2-Digit SIC code

a	Manufacturing – Nondurable (SIC 20-23, 26-29)
b	Manufacturing – Durable (SIC 24-25, 30-34, 37-39)
c	Manufacturing – Machinery (SIC 35-36)
d	Transportation, Motorfreight, Transportation and Warehousing, Utilities (SIC 40-42, 44-47, 49)
e	Communications (SIC 48)
f	Wholesale Trade (SIC 50-51)
g	Retail Trade and General Merchandise Stores (SIC 52-59)
h	Finance, Insurance, Real Estate (SIC 60-67)
i	Services except Business and Health (SIC 70, 72, 75-76, 78-79, 81-84, 86-89)
j	Business Services (SIC 73)
k	Health Services (SIC 80)
l	Other (SIC 01-02, 07-17)

Because of the greater importance of establishments with larger numbers of employees (only partially accounted for by SIZETYPE), the sample was allocated among the 288 strata in proportion to the total number of employees reported (on the FY2000 EEO-1 form) by establishments in the stratum. For example, the first stratum (SIZETYPE = 1, INDUSTRY = a, REGION = B) contained 199 establishments, which reported a total of 17,123 employees; that total number of employees represented 0.299% of the overall total number of employees, 5,717,421, so the initial target for the sample of 10,000 establishments was an allocation of 29.9 establishments in that stratum.

As the example illustrates, the basic calculations for allocating the sample generally did not yield an integer, whereas the actual sample size in each stratum must be an integer. More importantly, those calculations could yield an allocation that exceeded the number of establishments in the stratum. This condition arose for all strata with SIZETYPE = 4.

The sampling algorithm took these practical complications into account. It arrived at an integer allocation for a stratum by starting with a non-integer allocation and choosing randomly between the next-smaller integer and the next-larger integer according to probabilities that gave the non-integer allocation as the average. Then, where the allocation exceeded the available number of establishments, it selected (with certainty) all the establishments in the stratum. At that point, sampling for those "certainty strata" was complete. The algorithm determined the total number of establishments in the certainty strata, subtracted that figure from the overall target (10,000), and re-allocated the remainder among the non-certainty strata. If these new allocations exceeded the available number of establishments in any strata, the establishments in those strata were sampled with certainty, and the step was repeated. As soon as the re-allocation produced no new certainty strata, a final step ensured a sample size of at least 8 in each stratum, by combining strata that had adjacent values of REGION (and the same value of SIZETYPE and INDUSTRY). Only 14 strata required such collapsing, leaving a total of 276 strata. Because of the random rounding in the allocation procedure, the actual total sample size was 10,018 establishments. The actual sample was obtained by selecting a simple random sample of establishments from each of the 276 final strata. The table in Appendix A also gives the distribution of the sample over the strata.

Subsample for Reviews

For the study it was necessary to conduct reviews on establishments that responded to the 2002 EO Survey. The findings from those reviews provided the outcome measure for the study, presence or absence of systemic discrimination. Thus, within the sample of 10,018 establishments selected for the EO Survey, a subsample were selected as candidates for reviews. The subsample was selected in three parts, an initial sample of 3,300 and two supplementary samples (of 1,000 and 2,100, respectively), as experience with the reviews led to revisions in the initial assumptions. Thus, the total size of the subsample was 6,400.

The OFCCP expected to be able to conduct 2,250 reviews for the study. Initially, among the establishments receiving the EO Survey, 80% were expected to give a substantive response. Of those responses, 90% were expected to submit complete data. And among the reviews undertaken, the OFCCP expected that 95% would be completed in time for inclusion in the database to be used in the study. Thus, the initial size of the raw sample was $2,250/(0.80\times 0.90\times 0.95) = 3,300.$

The simplest approach for selecting the subsample would have applied systematic random sampling to a file containing the main sample in stratum order (using 10,018/3,300 = 3.0358 as the sampling interval). This approach, however, could not be applied to the entire sample in a single step, because it would have produced a subsample containing too many establishments

with the same parent company. In practice, most companies can handle up to about nine compliance reviews in a fiscal year without asking for special consideration (e.g., in scheduling), but for some companies a straightforward systematic random sample would have contained a much larger number of establishments. The most extreme two had 105 and 93 establishments, respectively, in the main sample and hence would have been expected to have over 30 establishments in the subsample. Thus the approach separated the main sample (of 10,018) into three parts and selected a portion of the subsample from each, in such a way that the subsample contained no more than nine establishments of any one parent company.

An initial step removed 28 establishments of an organization on which the OFCCP no longer conducts reviews, so that they would not be in the subsample.

The first part of the remaining sample consisted of only the 198 (=105 + 93) establishments of the two parent companies mentioned above, sorted by company and then by stratum. From it a selection interval of 11.647 (= 198/17) was used to select a systematic random sample. The resulting subsample contained 9 of the 105 establishments of one parent company and 8 of the 93 establishments of the other.

The second part of the sample consisted of the 912 establishments whose 24 parent companies had from 27 to 54 establishments in the main sample (in the frequency distribution of establishments by parent company, 54 followed 93). From the corresponding file, sorted by parent company and then by stratum, a selection interval of 6.0 (= 54/9) was used to select a systematic random sample of 152 establishments. In this subsample the number of establishments per company ranged from 4 to 9.

The third part of the sample consisted of the remaining 8,880 establishments (8,880 = 10,018 – 28 – 198 – 912). To complete the initial subsample, 3,131 (= 3,300 – 17 – 152) establishments were selected from it, using systematic random sampling with a selection interval of 2.8362 (= 8,800/3,131). In the resulting subsample no company had more than 9 establishments.

On the basis of partial response to the survey, in April 2003, OFCCP revised its estimate of the percentage of establishments that it expected to give a substantive response, from 80% to 62%. This change increased the size of the subsample, from 3,300 to 4,300. Thus the second part of the subsample consisted of 1,000 establishments, selected from the 5,749 (= 8,880 – 3,131) then remaining in the third part of the main sample (described above) by systematic random sampling with a selection interval of 5.7490 (= 5,749/1,000).

Subsequently, in August 2003, OFCCP estimated that the first two parts of the subsample would produce 1,600 to 1,800 reviews, that 57% of the establishments were giving a substantive response, and that, among establishments added to the subsample, 55% would have complete data on the EO Survey and be completed in time for inclusion in the database. Thus, in order to bring the total number of completed reviews back up to 2,250, a further supplemental subsample of $650/(0.57 \times 0.55) = 2{,}100$ establishments was drawn by systemic random sampling from the 4,749 (= 8,800 – 3,131 – 1,000) then remaining in the third part of the main sample. The selection interval was 2.261 (= 4,749/2,100).

Chapter 3. Data Collection

Survey Instrument

The data on which this study is based were collected on the OFCCP Equal Opportunity Survey of Federal Contractor Establishments (EO Survey, reproduced in Appendix B). The instrument has three parts:

- Part A – General information
- Part B – Personnel activity
  - Applicants
  - Hires
  - Promotions
  - Terminations
  - Full-time employees at end of year
- Part C – Annual monetary compensation and tenure
  - Annual monetary compensation
  - Average tenure

Both Part B and Part C ask for information on the nine EEO-1 categories (officials and managers, professionals, technicians, sales workers, office and clerical, craft workers, operatives, laborers, and service workers), broken down by gender and race and ethnicity. In Part B each area of personnel activity has a separate page that presents, for each EEO-1 category, a cross-classification of gender and the following categories of race and ethnicity:

- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or Other Pacific Islander
- White
- Hispanic or Latino (all races)
- Hispanic or Latino (White race only)
- Hispanic or Latino (all other races)

The page for applicants has an additional category, Race unknown, for situations in which "a resume or application that is screened is received without any racial or ethnic identification and no further contact is made with the applicant." Persons who identify with more than one racial category are to be counted only once. The apparent redundancy in the categories for Hispanic or Latino accommodated establishments whose recordkeeping systems did not yet distinguish Hispanics or Latinos as an ethnicity and identify Hispanics and Latinos by the five racial

categories. In completing Part B respondents had the option to use the full set of categories of race and ethnicity listed above (FORMAT 1) or (FORMAT 2) to follow the same format as in FORMAT 1 except:

- Record all actions pertaining to Hispanics or Latinos in the "Hispanic or Latino (all races)" columns.
- Leave the "Hispanic or Latino (White)" and "Hispanic or Latino (all other races)" columns blank.
- Record all actions pertaining to Asians, Hawaiians, and Other Pacific Islanders in the "Asian" columns.
- Leave the "Native Hawaiian or Other Pacific Islander" columns blank.

In Part C the information on race and ethnicity forms two categories:

- Non-minority – "someone of the White race who is not of Hispanic (or Latino) ethnicity"
- Minority – "all races other than White or someone of the White race who is of Hispanic (or Latino) ethnicity, or someone who has reported more than one race."

When combined with gender, these categories yield four groups of employees. For each combination of group and EEO-1 job category, Part C asks for information on

- Total annual monetary compensation for all employees
- Lowest annual monetary compensation of any single employee
- Highest annual monetary compensation of any single employee
- Average tenure with firm (in years and months)

The employees whose information is to be reported in Part C are precisely the full-time employees reported in Part B.

For the reporting period in Part B (and Part C), the respondent indicates the choice between the most recently concluded calendar year and the most recently concluded Affirmative Action Program year (not January 1 through December 31). The numbers of full-time employees are to be reported as of the last day of the chosen reporting period. The numbers of applicants, hires, promotions, and terminations reflect totals of events that occurred during that reporting period.

These reporting definitions are natural and reasonably convenient, but they pose challenges for analyses – in particular for developing predictor variables (Chapter 4) that are defined comparably for a wide range of establishments. For example, one might like to express the number of women hires (or, respectively, men, minorities, or non-minorities) as a simple

percentage of the corresponding number of applicants. In practice, however, some (or even all) of the hires may have applied during the previous reporting period. As long as the denominator is not zero (or, preferably, not small), the ratio of hires to applicants still yields a rate; but the resulting rates are not necessarily comparable between, say, females and males. Similarly, the number of full-time employees on the last day of the reporting period may be only a rough approximation of the number who should be considered in calculating rates of promotion and termination, especially if the establishment underwent substantial growth or shrinkage during that reporting period.

As an alternative to returning the hard-copy questionnaire, establishments could respond electronically by using a version of the EO Survey on a website. Those who chose that mode of response had available some additional flexibility: they could report their data by Affirmative Action Program Job Group ("a collection of jobs in an organization with similar content, wage rates, and opportunities"), in lieu of the nine EEO-1 job categories. For each AAP Job Group, however, they had to indicate the corresponding EEO-1 category. Also, those who chose to report by AAP Job Group had to use those Job Groups consistently throughout Part B and Part C.

Response on the 2002 EO Survey

The mailing of the EO Survey to the establishments in the sample, the processing of the responses, and the creation of the database were handled by Eastern Research Group under a contract with the Department of Labor.

The questionnaires were mailed in December 2002, and responses were supposed to be submitted by February 28, 2003. In practice, some responses took longer to arrive. At the end of March, 2,187 of the 10,018 establishments (nearly 22%) had not responded. Another 510 (5%) had been finalized as out of business, refused, or final return by the Post Office. Also, 175 had been returned by the Post Office and had been remailed. And 2,107 (21%) had asserted that they were not subject to the survey (i.e., the OFCCP lacked jurisdiction over them).

The remaining 5,039 (50%) had been entered into the database or were at various stages in the process of receipt and data entry.

On January 23, 2004 Abt Associates Inc. received access to the database containing the basic data (responses and other outcomes) from the 2002 EO Survey. A revised database became available on March 5, 2004. The database records the status of the EO Survey for each of the 10,018 establishments in the sample, according to an extensive list of status codes. Table 3.1 summarizes the final status codes by combining them into five disposition categories.

Table 3.1 Summary of Final Dispositions, 2002 EO Survey

	Number	Percent
Status "OK"	4,254	42.5
Other surveys with data	1,522	15.2
Nonrespondents	1,004	10.0
Asserted no jurisdiction	2,738	27.3
Out of business	500	5.0
Total	10,018	100.0

For reasons discussed below, the main analytic work for the study used only the data from the 4,254 surveys (42.5%) whose status was "OK"; that is, the establishment submitted a substantive response, and the data passed the prescribed edit conditions. Another 1,522 surveys (15.2%) had data, but their final status code indicated some limitation or problem; for the majority of these (949) an "edit condition report" had been generated, but the problem with the data had not been resolved. Various forms of nonresponse accounted for 1,004 establishments (10.0%). Finally, establishments that contested jurisdiction (2,738, 27.3%) and establishments that had gone out of business (500, 5.0%) were regarded as out of scope.

Preliminary analyses of the data, focusing primarily on the 4,174 surveys with final status "OK" in the initial version (January 23, 2004) of the database, raised some concerns about the quality of the responses. In one of the most striking examples, one establishment reported that it had 22 full-time minority female office and clerical workers and that their total annual monetary compensation was $46,775,457,059 (the corresponding lowest and highest monetary compensation were $16,224 and $31,928). That was, however, not entirely an isolated occurrence. A review of the compensation data, screening for situations in which the average compensation (calculated as the ratio of total annual monetary compensation from Part C to the number of full-time employees from Part B) was less than the lowest compensation or greater than the highest compensation, turned up eight additional instances in which the reported total annual compensation exceeded $1 billion.

A more systematic summary calculated, for each of the 36 combinations of female/male, minority/non-minority, and EEO-1 job category, the percentage of surveys that had average annual compensation greater than the highest annual compensation. That percentage ranged from 0.68% to 4.46%. Aggregating to the level of the survey, 753 (18.0%) of the 4,174 surveys with status "OK" had some problem with average compensation (either above the highest or below the lowest) for at least one of the 36 combinations. The presence of so many problems in surveys with status "OK" was rather surprising, but the edit checks performed when the data were incorporated in the database did not include any comparison of average annual compensation against lowest or highest annual compensation. The anomalies described above made the compensation data (which were not regarded as highly reliable) more difficult to work with; but, with suitable cleaning, they were still usable.

To a lesser extent the data on tenure also contained values that were at best suspect. For example, one establishment reported 9 full-time white male sales workers with an average tenure of 74 years. Another reported 11 full-time white male officials and managers with an average

tenure of 52 years. These were the only values of average tenure, however, that exceeded 50 years, and the bulk of values that exceeded 35 years were based on one or two employees. Thus, the two high values were set aside as outliers, and the rest of the tenure data were retained.

In Part B some data values seemed clearly inconsistent with the surrounding data. For example, one survey yielded, for female laborers, 0 applicants, 0 hires, 0 promotions, 339 terminations, and 1 full-time employee; its number of male terminations in that EEO-1 category was also anomalous. For female office and clerical workers, another survey showed 20 applicants, 11 hires, 4 promotions, 125 terminations, and 34 full-time employees. Yet another survey had, for male laborers, 787 applicants, 42 hires, 0 promotions, 23 terminations, and 5 full-time employees; and, though not so extreme, its data in other EEO-1 categories did not seem reasonable. Thus, before data from Part B could be used for predictor variables, it was necessary to devise and apply a rule (described in Chapter 4) that screened for problems in any of the components of Part B.

Another concern in Part B involved the reporting of employees' Hispanic ethnicity. As mentioned above, the directions for Part B seemed to indicate that respondents would report employees by Hispanic ethnicity (White race only versus all other races or all races) and separately report them by the five racial categories. This expectation was not borne out by the data. Some of the data were not consistent with regarding the employees reported as "Hispanic or Latino (White race only)" as a subset of the employees reported as "White." In one survey, for example, the number of full-time White employees was 2, and the corresponding number of full-time Hispanic White employees was 81. According to clarification received from OFCCP, respondents interpreted "White" according to the earlier definition "White not of Hispanic origin" and took a similar approach to the other racial categories. The data supported the conclusion that the other four racial categories (other than "White") were being interpreted as disjoint from "Hispanic or Latino (all other races)." Thus, it appeared reasonable, in various calculations, to form totals by adding data from all categories of race and ethnicity (or to form subtotals for minority or non-minority by adding data from the categories that correspond to those definitions).

Review of Establishments in the Subsample

For establishments in the review subsample (described in Chapter 2) that responded to the 2002 EO Survey, OFCCP proceeded with a compliance review.

Review Process

The procedures for such reviews are documented in detail in the Federal Contract Compliance Manual, which is available on the OFCCP website: http:\\www.dol.gov\esa\regs\compliance\ofccp\fccm\fccmanul.htm. Briefly, among the preparatory steps, the Compliance Officer conducting the review notifies the contractor establishment, via a Scheduling Letter, of the compliance review and requests the Affirmative Action Program (AAP) and supporting documentation. The Compliance Officer also sends inquiries to various agencies, to gather information on complaints filed against the contractor and other information pertinent to the review, and examines files from previous compliance actions (if any). The review covers at least the last full AAP year, and the Compliance Officer ordinarily attempts to complete the process within 60 days.

The desk audit of the AAP and related materials includes an evaluation of its analysis of the contractor's workforce, its analysis of the availability and utilization of minorities and women, and its goals for job groups that have been identified as underutilized. Other areas reviewed include personnel activity (e.g., hires, promotions, and terminations) and wage and salary data. The process may trigger requests for clarification and additional information. Some questions and problems may be resolved during the desk audit. In almost all instances, however, a full evaluation requires a further, onsite review.

The onsite phase of the compliance review aims primarily to investigate problem areas identified in the desk audit, to verify the contractor's implementation of its AAPs, and to begin to resolve violations. The Compliance Officer collects additional information by examining the contractor's files, making visual observations, and conducting interviews and discussions. The onsite review uses the information to examine compliance with a wide variety of requirements and to investigate areas of potential employment discrimination. At the end of the onsite review the Compliance Officer discusses the findings with the contractor in an exit interview and subsequently provides them in writing.

Dispositions of Reviews

The actual reviews for establishments in the review subsample began during the summer of 2003. On August 20, 2004 Abt Associates received a spreadsheet containing the results of the reviews. Table 3.2 summarizes the dispositions from the review process for the full subsample of 6,400 establishments. In view of the substantial numbers of establishments that did not respond (including those that challenged jurisdiction) to the 2002 EO Survey itself (Table 3.1), it is not surprising that reviews were not opened for 48% of the establishments in the subsample.

Table 3.2 Summary of Final Dispositions from Review Process, Total Subsample, 2002 EO Survey

	Number	Percent
Systemic discrimination	89	1.4
No systemic discrimination	2,601	40.6
Review started but not completed	22	0.3
Not reviewable	592	9.2
Review never opened	3,096	48.4
Total	6,400	100.0

Table 3.3 Summary of Dispositions from Review Process, Establishments Whose 2002 EO Survey Had Status "OK"

	Number	Percent
Systemic discrimination	67	2.2
No systemic discrimination	2,159	70.8
Review started but not completed	9	0.3
Not reviewable	378	12.4
Review never opened	435	14.3
Total	3,048	100.0

Because this study initially used data only from surveys that had status "OK," it is more informative to consider the dispositions for the 3,048 establishments in the subsample whose surveys were "OK" (Table 3.3). (Table B in Appendix E provides more detail on the relation between the establishments in Table 3.3 and those in Table 3.2. In particular, of the 22 [= 89 – 67] establishments in Table 3.2 that had systemic discrimination but whose surveys did not have status "OK," 16 had surveys with data and a status close enough to "OK" that they could be included [along with 299 establishments that had a finding of no systemic discrimination] in an augmented set of data, on which a substantial part of the analysis was rerun, as described in Appendix E.) The first two rows separate the 2,226 establishments with completed reviews according to whether the finding was systemic discrimination (n = 67) or no systemic discrimination (n = 2,159). These are the surveys on which the model building (Chapter 5) was based. A review may have been "never opened" or "started but not completed" for a variety of reasons. For example,

- The Regional Office, but not the National Office, had a record of a review within the previous two years (so that the establishment was not eligible for a review);
- After the survey was completed, events made it impossible to open a review (e.g., the establishment merged with another, closed, or burned down); and
- The completed survey was received after the last schedule of reviews was drawn up.

Some similar reasons and others accounted for establishments that were "not reviewable":

- The establishment had been reviewed within the previous two years;

- Events made a review impossible;
- The AAP included fewer than 50 individuals;
- The AAP combined the employees at the establishment with employees of another establishment; and
- The company asserted that it was not a Federal contractor, and the OFCCP was not able to locate a contract that demonstrated coverage.

Thus, some 73% of the review subset's surveys with status "OK" were available for model building.

The last column of the table in Appendix A shows the distribution of those 2,226 surveys over the 276 strata. The number of those surveys per stratum was generally in reasonable agreement with what one would expect from the corresponding numbers of establishments in the sample of 10,018, establishments in the review subsample, and surveys with status "OK." Only 3 of the 276 strata were not represented among the 2,226 surveys, and those three strata had small sample sizes in the subsample and small numbers of "OK" surveys.

On the whole, the rates of response for the survey sample and the review subsample were reasonably close to the assumptions made in adjusting the size of the review subsample (discussed at the end of Chapter 2). Some 58% of surveys were returned with data (Table 3.1), and 52% (=2,226/4,254) of the establishments with "OK" surveys had a completed compliance review.

Chapter 4. Predictor Variables

The building blocks for models aimed at targeting establishments that are more likely to be engaged in systemic discrimination are a variety of potential predictor variables, derived from data in Part B and Part C of the 2002 EO Survey. The process of developing the predictor variables for this study took into account the behavior of the data from that survey, as well as previous work on discrimination and its correlates and predictors. Among previous efforts the one most directly related to the present study is the analysis of the data from the 2000 Pilot Test of the EO Survey. The reports by Bendick and Miller (2000) and Bendick et al. (2000) discuss predictors from that pilot test. This chapter describes the predictors developed in this study. The SAS coding, delivered separately, gives full details.

In order to use as much data as possible, the analyses leading to these predictors were based on the data from the 4,254 surveys whose final status was "OK" (discussed in Chapter 3). If a respondent used AAP Job Groups instead of the nine EEO-1 categories, the data were aggregated to the level of the EEO-1 category.

Nearly all the predictor variables fall into two broad groups. One group attempts to measure the treatment of females relative to males. The other group, in a parallel fashion, compares the treatment of minority persons with that of non-minority persons. Within the two groups the variables separate into basic variables and comparative variables. Each basic variable is derived only from the data of the individual establishment. A corresponding comparative variable reflects the extent to which the individual establishment departs from the establishments in its comparison group, defined by industry and geography. The 12 industry groupings (Table 2.1) and the 9 Census Divisions yielded a total of 83 comparison groups (Appendix C), after adjacent Divisions were combined. The target in combining was a minimum of 25 establishments in each comparison group. The process stopped, however, when all Census Divisions within a Census Region had been combined. Ten of the comparison groups contained fewer than 25 establishments; but only three contained fewer than 20, and none had fewer than 16.

Part B

As described in Chapter 3, Part B requests data on applicants, hires, promotions, terminations, and full-time employees. These data were the basis for a variety of predictor variables. The general approach was to calculate a component for each EEO-1 category and then summarize those components to produce a value for the establishment. One summary, also used earlier, was the average. To allow for the possibility that an establishment's good performance on some EEO-1 categories could, in the average, make up for its poor performance on one category, a further summary selected the component from the most extreme category (usually the minimum).

The components for the EEO-1 categories compared two rates – for example, the ratio of female hires to female applicants and the ratio of male hires to male applicants. They expressed the comparison in two ways: the ratio of one rate to the other and the difference between the two rates. Ratios and differences are both customary forms of comparison. In the data from Part B,

however, the ratio could magnify small differences unduly. For example, the pairs 0.1 and 0.2, 0.2 and 0.4, and 0.4 and 0.8 all yield a ratio of 0.5, but the difference between 0.1 and 0.2 is often more likely to be the result of chance fluctuations than the difference between 0.4 and 0.8.

Hiring

It is natural to form rates for females and males (and for minority persons and non-minority persons) by dividing the number of hires during the year by the number of applicants during the year. One set of predictor variables was based on these rates. This definition of rate, however, has a drawback: the persons hired during the year need not be a subset of the persons who submitted applications during the year. Those who were hired early in the year may have applied during the previous year, and those who applied late in the year may still have been under consideration at the end of the year. Thus, a ratio of hires to applicants would generally not have the customary statistical properties of a proportion. In particular, it need not be between 0 and 1. The data from the 2002 EO Survey contain a number of instances in which the number of hires exceeds the number of applicants and the data from the establishment seem reasonable.

Although they do not avoid this ambiguity, an alternative set of predictor variables used the ratio of the number of hires during the year to the number of full-time employees at the end of the year. In using the number of full-time employees as the denominator, however, one is not measuring the same thing as when one uses the number of applicants. If women (or minorities) are underrepresented among full-time employees and among hires, this measure may not reflect the deficiency. In the absence of such a problem, however, the number of full-time employees should generally be more stable than the number of applicants.

Comparative Variables

For hiring, as for the other areas of activity, the basis for comparison in the comparative variables is the median, over the establishments in each comparison group, of the basic component values for an EEO-1 category. Through the median each establishment is compared with all the establishments in its comparison group, including itself. Thus, each comparison group has a median for each of the nine EEO-1 categories. For example, the median ratio of female hire rate to male hire rate for professionals is the median of those (non-missing) ratios from the establishments in the comparison group.

For an individual establishment, the calculations for the comparative ratio variable take the ratio of its basic component value for each EEO-1 category to the corresponding comparison-group median. Any of the resulting ratios that exceed 1 are set to 1 to limit their effect in offsetting low values in other categories. Finally, those adjusted ratios are averaged over the EEO-1 categories to produce the value of the comparative ratio variable for the establishment.

The calculations for the comparative difference variable follow the same steps, using differences instead of ratios and applying 0 as the ceiling instead of 1.

If, as occasionally happened in the 2002 data, a comparison group contains only one non-missing value of the basic component for an EEO-1 category, the median for that category was set to missing, so as to avoid an artificial ratio of 1 or difference of 0 for the one establishment to which the non-missing value belonged. The SAS coding documents the treatment of other special cases.

Promotions and Terminations

For rates of promotion and rates of termination the only quantity readily available to serve as a denominator is the number of full-time employees at the end of the year. Thus, this study used that denominator. The numerators for the rates of promotion and the rates of termination were the corresponding numbers of promotions and terminations, respectively, during the year. The definitions of the predictor variables for promotions and the predictor variables for terminations were parallel in all respects.

Both the rates of promotion and the rates of termination have a potential limitation analogous to that discussed above for rates of hiring. The number of full-time employees at the end of the year may not be a good indicator of the number of employees, at various times during the year, who might be candidates for promotion or at risk of termination, especially if the number of full-time employees increased or decreased substantially over the course of the year. Ideally, however, such patterns should have the same effect on women, men, minorities, and non-minorities.

Treatment of Anomalous Data

Examination of the data from Part B revealed a number of instances in which the number of promotions or the number of terminations was substantially higher than the number of full-time employees, or the number of hires was substantially higher than the number of applicants or the number of full-time employees. Some of these instances could conceivably reflect establishments that had high rates of turnover, but further analysis suggested that either the data were simply bad or the number of full-time employees was small enough that various ratios were unstable. To reduce the adverse impact of such situations, a screening rule was applied to the data at the level of the EEO-1 category. If, for either females or males, the ratio of hires to applicants exceeded 2 or the ratio of hires to full-time employees exceeded 2 or the ratio of promotions to full-time employees exceeded 2 or the ratio of terminations to full-time employees exceeded 2, then all eight of the rates involved were set to missing. A parallel rule was applied to the corresponding eight rates for minorities and non-minorities.

Full-time Employees

The numbers of full-time employees yielded a number of predictor variables, derived from the percentage distributions of females and males (and minorities and non-minorities) over the nine EEO-1 categories. For example, let f_i denote the proportion of female full-time employees in

category i, and let m_i denote the corresponding proportion for male full-time employees, so that $f_1 + f_2 + ... + f_9 = 1$ and $m_1 + m_2 + ... + m_9 = 1$. Also, let d_i denote the absolute value of the difference between f_i and m_i: $d_i = |f_i - m_i|$. Then one predictor was an index of female "occupational segregation" that weights the nine EEO-1 categories as prescribed in the EEDS manual (also used in the analysis of the 2000 Pilot Test):

$$3 - \frac{1}{2}\left(3d_1 + 2.21d_2 + 1.83d_3 + 1.61d_4 + d_5 + 1.79d_6 + 1.45d_7 + 1.18d_8 + d_9\right).$$

The theoretical maximum of this expression is 3, corresponding to $d_1 = d_2 = ... = d_9 = 0$ (i.e., no occupational segregation), and occupational segregation increases as the value decreases. A corresponding index of minority occupational segregation was derived, in parallel fashion, from the percentage distributions of minority and non-minority full-time employees over the EEO-1 categories.

In addition, each of the nine d_i for females versus males and each of the nine d_i for minorities versus non-minorities was made into a separate predictor variable. This strategy made the individual differences available, without averaging against one another.

The female occupational segregation variable and the minority occupational segregation variable both had two corresponding comparative variables: one based on the ratio to the median in the comparison group (at the level of the establishment) and the other based on the difference from that median.

The data on full-time employees provided five additional predictor variables:

- The number of non-empty EEO-1 categories that contained 0 female employees;
- The number of non-empty EEO-1 categories that contained 0 minority employees;
- An indicator (0 or 1) of whether any EEO-1 category with more than 10 male employees contained 0 female employees;
- An indicator of whether any EEO-1 category with more than 10 non-minority employees contained 0 minority employees;
- An indicator of whether the total number of full-time employees was greater than 200.

Part C

As mentioned in Chapter 3, Part C requests information on annual monetary compensation (total for all full-time employees, lowest for any single employee, and highest for any single employee) and average tenure. Each of these areas produced a number of predictor variables. As in the development of predictors from Part B data, the general approach was to calculate a component

for each EEO-1 category and then summarize those components (by taking the average and, usually, the minimum) to produce a value for the establishment. The components were generally ratios of female to male or minority to non-minority; for data on compensation and tenure, which often are farther from zero and have substantially wider ranges of values than the rates calculated in Part B, ratios are usually more appropriate for comparisons than differences.

Similarly, the procedure for developing comparative variables in Part C closely paralleled that in Part B (described above).

Compensation

Within each EEO-1 category the component variable for compensation among females (or among males or minorities or non-minorities, respectively) was the Adjusted Average Wage (AAW). When *n*, the number of female full-time employees, is at least 3, the adjustment subtracts the lowest compensation and the highest compensation from the total compensation and divides the result by $n-2$. This approach to estimating compensation, previously used in the analysis of the data from the 2000 Pilot Test, is in the spirit of the statistical techniques known as trimming. When *n* is 1 or 2, the only reasonable estimate divides total compensation by *n*.

In all these cases the data are first required to satisfy basic consistency checks, such as whether the average annual compensation (total compensation divided by *n*) is greater than or equal to the lowest annual compensation and less than or equal to the highest annual compensation. Otherwise, AAW is set to missing.

In preparation for these calculations, various data must be combined because, for example, Part C collects data separately for minority females and non-minority females. Total compensation for females can be obtained by adding the values for minority females and non-minority females. Conveniently, the highest single compensation for females is the larger of the two values for minority females and non-minority females, and similarly for the lowest single compensation. Also, the total number of female full-time employees can be obtained from the more-detailed breakdown in Part B.

Parallel calculations yielded values of male AAW, minority AAW, and non-minority AAW for the EEO-1 category. As mentioned above, the respective components were the ratio of female AAW to male AAW and the ratio of minority AAW to non-minority AAW. In some instances these AAW ratios seemed unreliable. The majority of questionable ratios were based on a very small number of employees in either the numerator or the denominator. Thus, an AAW ratio was set to missing unless the AAWs in both its numerator and its denominator were based on at least three employees. For some establishments these missing AAW ratios caused the average over non-missing values to be missing, but an average of unreliable values would not have been reliable.

Tenure

The data on average tenure of full-time employees did not require any special preparation, other than setting aside two implausibly large values, as mentioned in Chapter 3. Once the tenure ratio for females to males and the tenure ratio for minorities to non-minorities had been calculated, however, their distributions (within EEO-1 category, over establishments) showed substantial skewness toward large values. A number of the more-extreme values came from ratios in which either the numerator or the denominator was based on a small number of employees (a source of instability). Thus, the ratio for females to males and the ratio for minorities to non-minorities were set to missing unless the values of average tenure in the corresponding numerator and denominator were both based on at least three employees. The remaining patterns suggested that the averages (over EEO-1 categories within establishment) might have better statistical behavior if the individual ratios were transformed to a different scale. This process, applying a suitable mathematical function to each data value, is frequently used in data analysis. For the tenure ratios two transformations were considered, the logarithm and the reciprocal. In terms of the numerator and denominator in the ratio, say x and y, the effects of these transformations are as follows:

$$\log(x/y) = \log(x) - \log(y)$$
$$\text{recip}(x/y) = y/x.$$

Both of these transformations made the distributions within EEO-1 category more nearly symmetric, but neither was clearly preferable to the other. Thus both were used as the basis for predictor variables, in addition to the untransformed tenure ratios.

The use of the two transformations had implications for certain predictor variables. Because, as shown above, the logarithm of a ratio is a difference, the comparative variables for the logarithmic scale used differences instead of ratios. And, because the reciprocal reverses the ordering on the numeric scale (e.g., 2<3 but 1/2 >1/3), the "minimum" variables for the reciprocal scale actually took the maximum over the EEO-1 categories.

Chapter 5. Model Building

The process of building models that attempt to target systemic discrimination brings together the predictor variables described in Chapter 4 and the finding (from the compliance review) of presence or absence of systemic discrimination, discussed in Chapter 3. For this dichotomous outcome the statistical models considered in this study are forms of logistic regression. The basic approach in such models is to express the probability of an event (here, a finding of systemic discrimination) as a combination of predictor variables, transformed by a mathematical function that gives values between 0 and 1. The subject of logistic regression has an extensive literature. The book by Hosmer and Lemeshow (2000) presents the mathematical formulation, discusses and applies a variety of statistical techniques, and provides access to related literature.

This chapter focuses on selecting suitable predictor variables and evaluating the resulting model. The major steps involve examining the data, fitting single-variable models, formulating multiple-variable models, and assessing the adequacy of the final model. SAS (versions 8 and 9.1) was used for data management and statistical analysis.

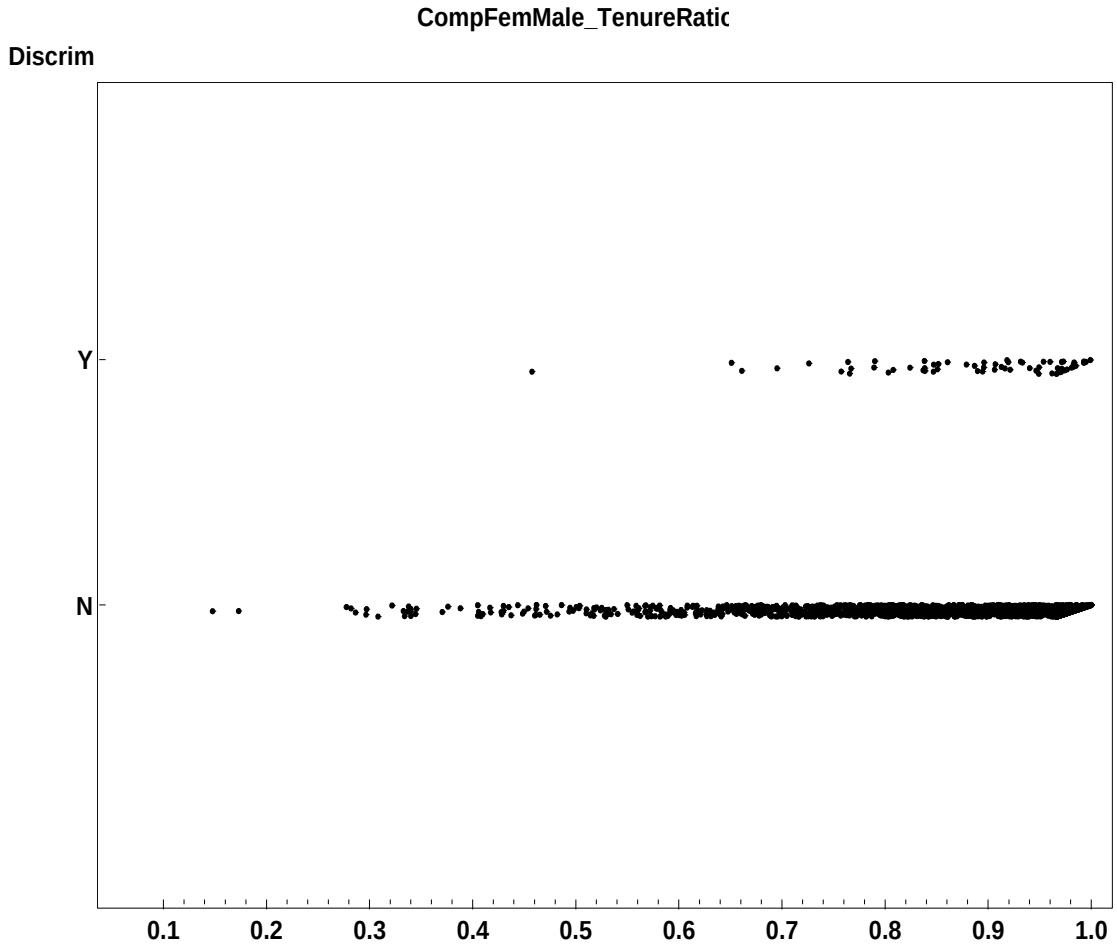
Single-variable Analyses

The first phase of analysis examined the relation between systemic discrimination (*SD*) and each of the predictor variables, via numerical summaries and graphical displays. If the predictor was dichotomous (e.g., Indicator_GT200) or had only a few discrete values, a cross-tabulation provided the essential detail. For predictors that had a large number of distinct values (and hence could be regarded as "continuous"), the primary summaries were sample percentiles, starting with the median and quartiles and moving outward as needed toward the minimum and maximum data values. (Such percentile-based detail is generally more informative than the mean and standard deviation.) The basic graphical display for such a continuous predictor plotted *SD* (*N* or *Y*, 0 or 1) against the value of the predictor.

Most predictors showed little relation to *SD*. The distribution of their values among establishments with *SD = Y* closely resembled that for establishments with *SD = N*. The two were centered at nearly the same value (e.g., the two medians were nearly equal), and they spread out in nearly the same way (sometimes with less spread in the *Y* distribution than in the *N).*

Some predictors showed a tendency for the *Y* values to concentrate near one end of the range of the *N* values. (None, however, came close to the type of pattern that would characterize a perfect predictor: absence of overlap between the distribution for *SD = Y* and that for *SD = N*.) Figure 5.1 illustrates this tendency in the predictor CompFemMale_TenureRatio. The much larger number of establishments with *SD = N*, however, makes it difficult to see whether a trend is present.

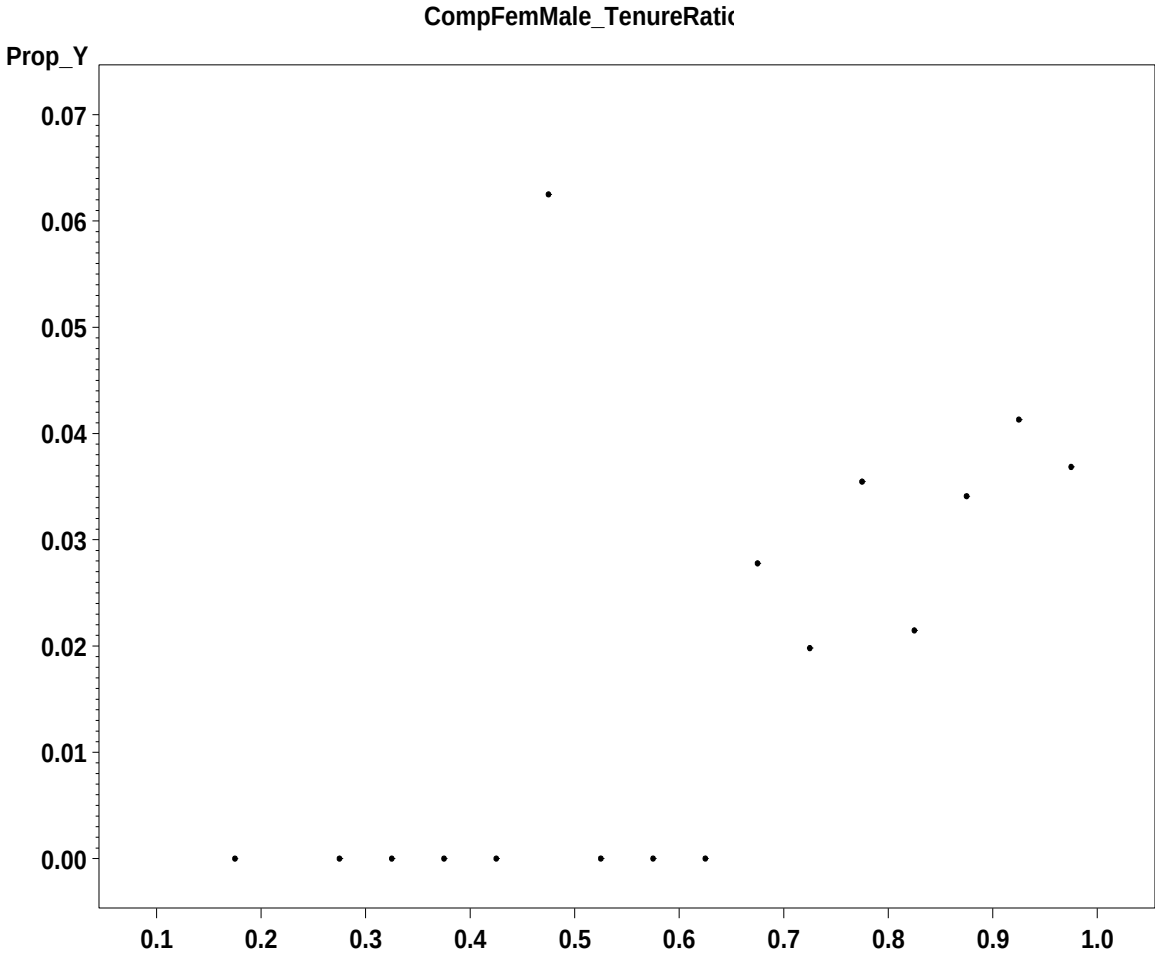
Figure 5.1 Scatterplot of Presence (Y) or Absence (N) of Systemic Discrimination against the Predictor CompFemMale_TenureRatio



An alternative display (a form of "smoothed scatterplot") is better suited to this task. One slices the horizontal axis into intervals, calculates the proportion of establishments with *SD = Y* in each interval, and then plots that proportion against the midpoint of the interval (i.e., one point per interval). Figure 5.2 shows such a plot for CompFemMale_TenureRatio, using intervals of width 0.05. Except for one establishment at 0.47, the *Y* distribution has no values below 0.65. Above this level the percentage with *SD = Y* increases in a jagged pattern. Though this pattern in the data is clear, it may be unexpected. As Chapter 4 explains, a comparative variable is derived by dividing the establishment's basic value (here, the ratio of average tenure for females to average tenure for males) for each EEO-1 category by the corresponding comparison-group median, capping the result at 1, and averaging over the EEO-1 categories. Thus, establishments with values of CompFemMale_TenureRatio close to 1 consistently have high values of the female-to-male tenure ratio, relative to their comparison group.

A second example illustrates a predictor variable whose values among establishments with *SD = Y* are concentrated toward the left end of the range. Figure 5.3 shows the ordinary scatterplot for

Figure 5.2. Smoothed Scatterplot, Showing the Proportion of Establishments with *SD = Y* in Intervals of CompFemMale_TenureRatio



MinWhite_TenureRatio. Except for two values at 1.7, the values for *SD = Y* range from 0.2 to 1.3, whereas the values for *SD = N* range from 0.04 to 4.2 (with two high outliers). Using intervals of width 0.5 yields the smoothed scatterplot in Figure 5.4, which shows a clear downward trend over the part of the range containing the data for *SD = Y*.

A further, more analytic, phase considered each predictor variable separately in a single-variable logistic regression model. At this stage in the model-building process one is not looking for statistical significance at the customary 0.05 level. Instead the aim is to identify predictor variables that have at least some association with the outcome variable. This study defined "some association" as a p-value less than 0.25 for the variable's coefficient in its logistic regression model. This broader search takes into account the possibility that several predictors, each having only "some association" with the outcome variable, may combine to make a strong contribution in a multiple-variable model.

Table 5.1 lists the 22 predictor variables whose p-value in their single-variable logistic regression was less than 0.25. Part B and Part C are both well represented (with 13 and 9 variables, respectively). Similarly, the list includes roughly equal numbers of female predictors and minority predictors. Within Part B promotions and full-time employees are the sources most often drawn upon (the latter mainly via the individual differences that go into the measures of occupational segregation). Interestingly, hires show up only in relation to full-time employees (rather than applicants), and none of the predictors are derived from terminations. The predictor with the most extreme statistical significance was the indicator of whether the establishment had more than 200 full-time employees ($p < .0001$). Within Part C variables derived from tenure substantially outnumber those derived from compensation.

Multiple-variable Analyses

The combined contributions of the 22 predictors listed in Table 5.1 were examined by including all of those variables in a multiple-variable logistic regression model. In that model only two variables, Indicator_GT200 and CompFemMale_TenureRRecip, had p-values smaller than .05 (.0015 and .0388, respectively). This result suggested that it would be appropriate to search for an intermediate set of predictors, containing substantially fewer than all 22 but more than two. Thus, stepwise logistic regression (which, at each iteration, considers whether any variable should be added to the model and then considers whether any variables in the model should be removed) was used, starting from the list of 22 predictor variables. The resulting model contained four predictors plus an intercept term, listed in Table 5.2.

Figure 5.3 Scatterplot of Presence (Y) or Absence (N) of Systemic Discrimination against the Predictor MinWhite_TenureRatio

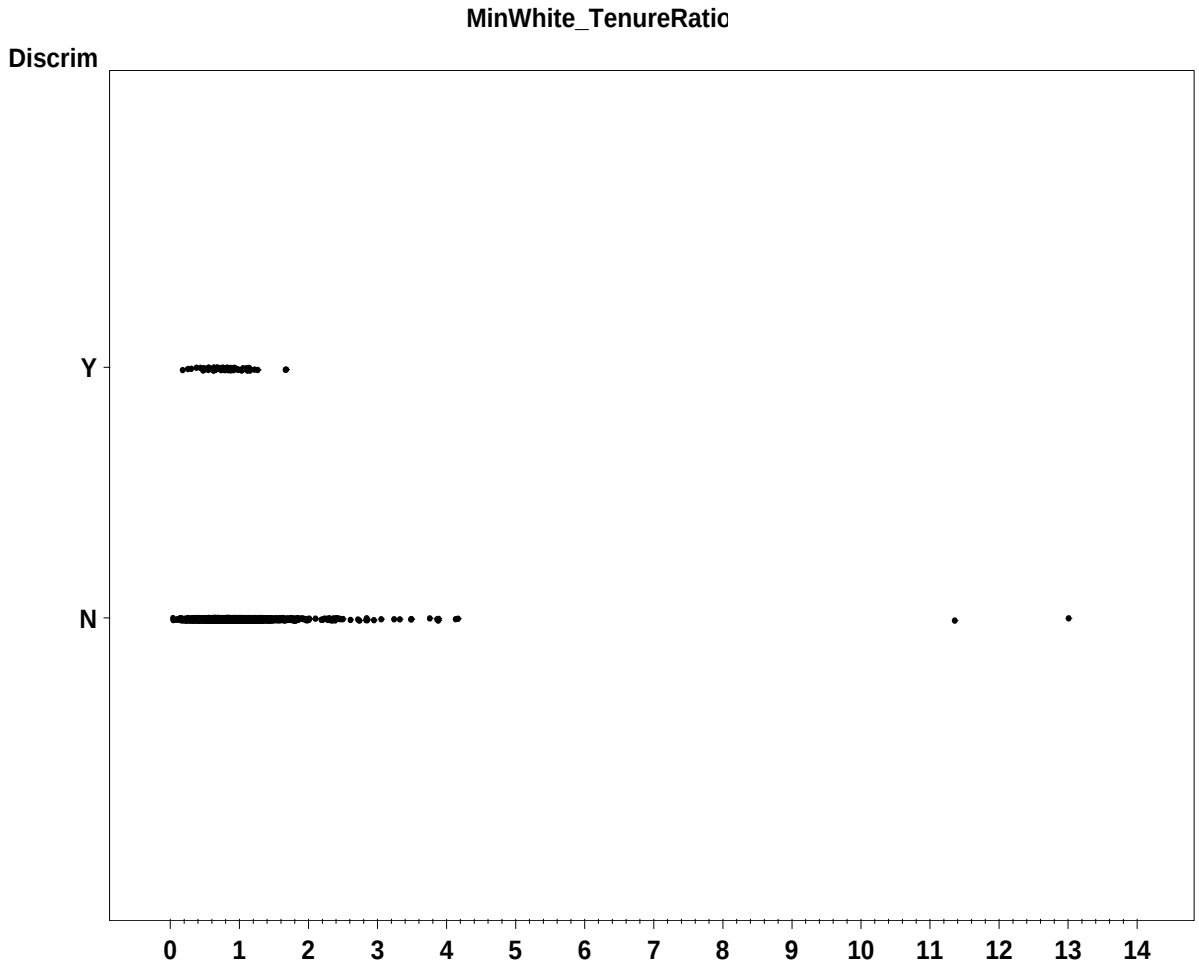


Figure 5.4 Smoothed Scatterplot, Showing the Proportion of Establishments with *SD = Y* in Intervals of MinWhite_TenureRatio

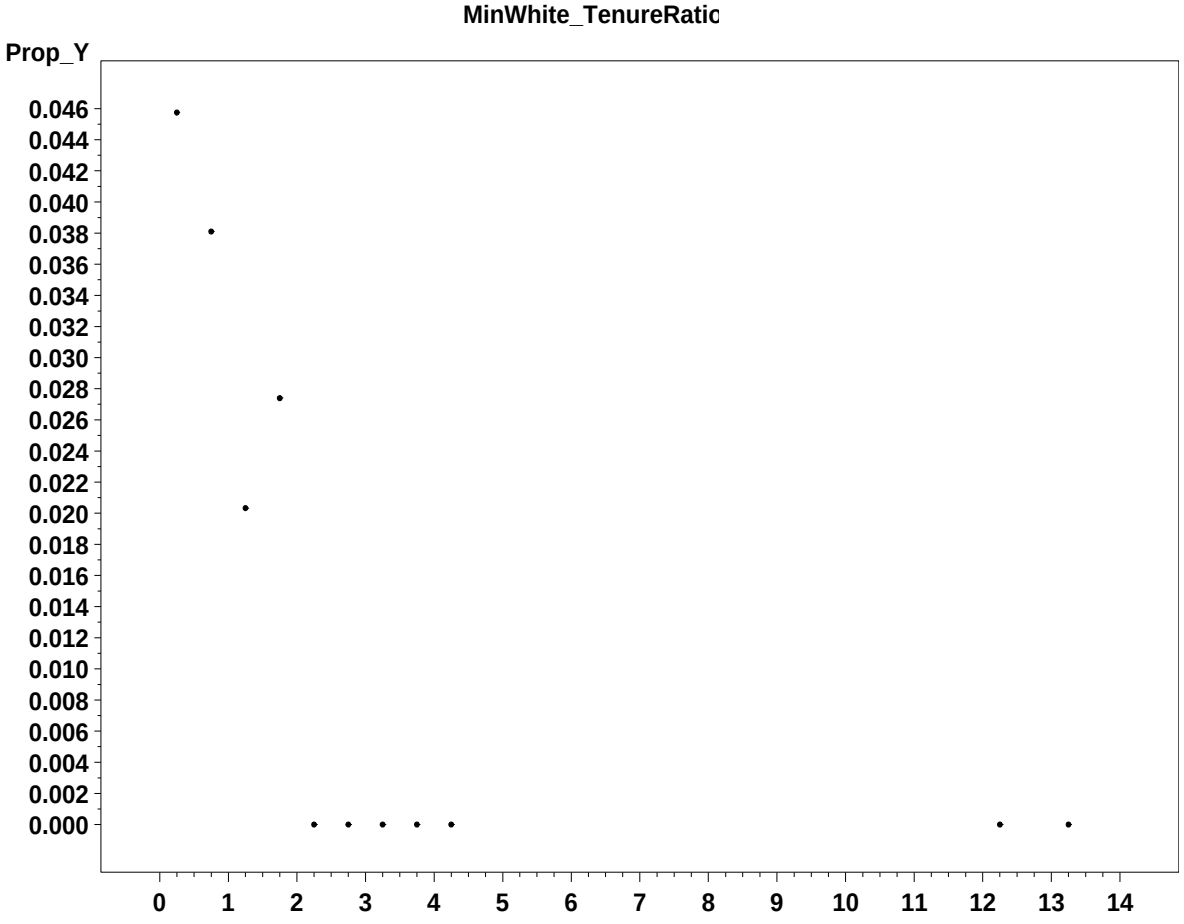


Table 5.1 Predictor Variables That Had $p < .25$ in Single-variable Logistic Regression Models, and Those p-values.

Predictor Variable	P-value
Part B	
FemMale_Ratio_Prom_Mean	.2366
FemMale_Diff_Prom_Comp_Mean	.2361
FemMale_Ratio_Hire_FT_Min	.2293
FemMale_Ratio_Prom_Min	.1772
FemMale_Diffi2	.0044
FemMale_Diffi3	.0006
Indicator_GT200	<.0001
MinWhite_Diff_Hire_FT_Comp_Mean	.0362
MinWhite_Diff_Prom_Comp_Mean	.2234
MinWhite_Diff_Prom_Comp_Min	.1947
MinWhite_Diffi2	.0197
MinWhite_Diffi3	.0020
MinWhite_Diffi5	.0938
Part C	
CompFemMale_AAWRatio	.1951
CompMinWhite_AAWRatioMinimum	.1886
FemMale_TenureRatioRecip	.1665
MinWhite_TenureRatio	.0027
MinWhite_TenureRatioMinimum	.0028
CompFemMale_TenureRatio	.0404
CompFemMale_TenureRLog	.0427
CompFemMale_TenureRRecip	.1806
CompMinWhite_TenureRLogMinimum	.0590

Table 5.2 Initial Model Obtained from Stepwise Logistic Regression (based on 1,500 of 2,226 observations)

Variable	Coefficient	Standard Error	P-value
Intercept	−6.186	1.495	<.0001
Indicator_GT200	1.776	0.524	.0007
MinWhite_TenureRatio	−1.690	0.602	.0050
FemMale_Diffi3	−8.550	3.613	.0180
CompFemMale_TenureRatio	3.058	1.532	.0458

For the stepwise search (as for the initial multiple-variable model), the number of observations used is determined by the requirement that each observation have non-missing values on all 22 predictor variables in the list. In this instance that subset of the data contained 1,500 of the 2,226 observations. A customary next step refits the selected model, using all observations that had no missing values on any of the four predictor variables in the model. Table 5.3 shows the refitted model, which used 1,891 observations (a gain of 391). As one would expect with the inclusion of a substantial number of additional observations, the coefficients changed somewhat, and their standard errors decreased. The largest change was in the coefficient of Indicator_GT200: from 1.776 to 1.228, a decrease of slightly more than its standard error in the initial model. Also, the p-value for CompFemMale_TenureRatio rose to .0553. The decision on whether to retain that variable in the model was deferred until after some further analysis.

The discussion of the scatterplot in Figure 5.3 noted the presence of two high outliers among the values for *SD = N*, and one low outlier is apparent among the values for *SD = Y* in Figure 5.1. The scatterplot for FemMale_Diffi3 (not shown) contained three outliers: one high value with *SD = Y* and two high values with *SD = N*. It is often advisable to set such data aside and refit the model, so as to avoid possibly adverse impacts from such a small fraction of the data and also to assess the extent of those impacts. Rerunning the stepwise analysis confirmed that the presence of the data from those six establishments did not affect the set of predictor variables selected. Refitting the model in Tables 5.2 and 5.3 produced the results in Table 5.4. Removing 6 of the 2,226 observations actually reduced the number of observations used by only 3, from 1,891 to 1,888, because the other three observations had missing values that had already prevented their use. Compared with Table 5.3, the changes in the coefficients were in the direction that one would expect from the position of the outliers in the scatterplots. Only the change in the coefficient of FemMale_Diffi3 was particularly large, and it was only slightly more than one standard error. The p-values for the four predictor variables were all considerably below .05. Thus, the final logistic regression model has the coefficients listed in Table 5.4. The evaluation and use of that model are discussed further in the next subsection of this chapter. (The analysis of the augmented data, summarized in Appendix E, led to the same four-variable model, fitted to 2,153 observations. Its coefficients, shown in Table C, are quite similar to those in Table 5.4.)

Table 5.3 Refitted Model with Four Predictor Variables (based on 1,891 of 2,226 observations)

Variable	Coefficient	Standard Error	P-value
Intercept	−4.928	1.233	<.0001
Indicator_GT200	1.228	0.337	.0003
MinWhite_TenureRatio	−1.386	0.493	.0049
FemMale_Diffi3	−7.154	3.031	.0183
CompFemMale_TenureRatio	2.424	1.265	.0553

Table 5.4 Refitted Model with Four Predictor Variables after Setting Aside 6 of the 2,226 Observations (1,888 observations were used)

Variable	Coefficient	Standard Error	P-value
Intercept	−5.318	1.312	<.0001
Indicator_GT200	1.193	0.339	.0004
MinWhite_TenureRatio	−1.492	0.506	.0032
FemMale_Diffi3	−10.685	3.726	.0041
CompFemMale_TenureRatio	3.040	1.345	.0239

From its review of the draft report, OFCCP expressed concern that the models (e.g., in Tables 5.3 and 5.4) did not include any predictors derived from compensation and asked whether that outcome might be the result of missing values on those predictors. Appendix E examines this question. The explanation does not seem to lie in the numbers of missing values.

Throughout this chapter the data are unweighted (i.e., all observations have equal weight). Appendix E describes the development of weights that incorporate a base weight from the selection of the main sample (Chapter 2), a factor that reflects the sampling fraction in the selection of the review subsample, and an adjustment for nonresponse on the compliance review. When the weights were used to refit the four-variable logistic regression model to the augmented data, the coefficients and standard errors (in Table E) differed somewhat from those in Table C, but not dramatically. Appendix E recommends the use of the results from the unweighted model.

When possible, data that have been set aside should be investigated further, to uncover the reasons for their anomalous behavior. In the present instance the detailed data from Part B or Part C for the six establishments provided clear explanations. In Figure 5.3 two establishments had strikingly high values of MinWhite_TenureRatio. The higher of these, 13.0, was the average of the ratios from two EEO-1 categories, 23.3 and 2.7; the first of these ratios was based on 3 minority employees with an average tenure of 163 months and 6 non-minority employees with an average tenure of 7 months, and the second was based on 3 minority employees with an average tenure of 307 months and 42 non-minority employees with an average tenure of 112 months. Even if all four values of average tenure were accurate, the small numbers of employees for three of them would justify limiting the impact of the establishment on the analysis.

The lower of the two extreme values of MinWhite_TenureRatio, 11.4, was the average of the ratios from three EEO-1 categories, one of which was 32.2, derived from 41 minority employees with an average tenure of 193 months and 3 non-minority employees with an average tenure of 6 months. A small number of employees again provides a reason for caution.

Small numbers of employees also played a role in the two high values of FemMale_Diffi3 with *SD = N*. In the establishment with the higher value, 0.89, 47 of the 53 male full-time employees were technicians (Category 3), and both of the 2 female full-time employees were office and clerical workers. In the establishment with the lower of the two extreme values, 0.70, 52 of the 74 male employees were technicians, and all of the 3 female employees were officials and

managers. The unusually high value with *SD = Y*, 0.36, came from an establishment in which 69 of the 190 male employees and 0 of the 51 female employees were technicians, and 65 males and 42 females were service workers (thus, the establishment's value of FemMale_Diffi9 was 0.48). In this instance the data suggest a substantial degree of occupational segregation. On FemMale_Diffi3, however, the establishment is a clear outlier (the remaining values with *SD = Y* ranged from 0 to 0.12), and retaining its data would have reduced the predictive ability of the model.

In Figure 5.1 the low value of CompFemMale_TenureRatio with *SD = Y*, 0.474, was the average of the comparative ratios from two EEO-1 categories, 0.395 and 0.553. For the first of these the establishment's ratio of tenure among female to male employees was 0.347, versus 0.878 in its comparison group; and its ratio for the second category was 0.65, versus 1.176 in the comparison group. The only small number of employees, however, was six female employees in the first category. Again, retaining this establishment's data would have reduced the predictive ability of the model.

Evaulating the Final Model

This subsection focuses on several aspects of the final model, including how well it fits the data and its predictive ability. For convenience Table 5.5 restates the model, whose coefficients are given in Table 5.4. To calculate the "predicted logit" for an establishment, one substitutes its values on the four predictor variables into the formula. Using $\hat{y}$ as an abbreviated notation for the predicted logit, the establishment's predicted probability of systemic discrimination is given by

$$Predicted\ probability = \frac{e^{\hat{y}}}{1+e^{\hat{y}}}$$

Over the 1,888 establishments whose data were used in fitting the model, the predicted logit ranged from –11.32 to –1.16, and the predicted probability ranged from 1.21×10^{-5} to 0.238. For comparison the overall rate of systemic discrimination among the 1,888 establishments was 63/1,888 = 0.033.

Table 5.5. Final Logistic Regression Model for Predicting the Presence of Systemic Discrimination (the formula gives values in the logit scale)

Predicted logit =
1.193 x Indicator_GT200
–1.492 x MinWhite_TenureRatio
–10.685 x FemMale_Diffi3
+3.040 x CompFemMale_Tenure Ratio
–5.318

One common question involves the degree of agreement between the predicted probabilities and the observed presence or absence of *SD*. An appropriate statistical test, described by Hosmer and Lemeshow (2000, Section 5.2.2), divides the establishments into ten groups according to their predicted probability, each group containing essentially the same number of establishments. The predicted probabilities within each group are averaged, and the average is used to estimate the expected number of establishments with *SD* within the group. The difference between the observed number and that expected number is squared, and a suitable function of the ten squared differences is referred to the chi-squared distribution on 8 degrees of freedom. Table 5.6 shows the result of applying the Hosmer-Lemeshow test to the predicted probabilities from the final model. The value of the test statistic (6.923) and its p-value (.545) indicate that model fits reasonably well. More-detailed evidence comes from comparing the observed and expected frequencies of *SD = Y* in the ten groups; agreement is good in each group. (The presence of numerous small expected frequencies is cause for some concern over the adequacy of the chi-squared approximation. The usual remedy would combine adjacent groups, say Groups 1–3 and Groups 4–5, and reduce the number of degrees of freedom accordingly. In these data, however, the effect on the qualitative result would not be large.)

Table 5.6 Observed and Expected Frequencies of Systemic Discrimination for the Ten Groups in the Hosmer-Lemeshow Test, in Order of Increasing Predicted Probability

	SD = Y		***SD = N***		
Group	**Obs.**	**Exp.**	**Obs.**	**Exp.**	**Total**
1	0	0.29	189	188.71	189
2	3	1.13	186	187.87	189
3	1	2.07	188	186.93	189
4	3	3.08	186	185.92	189
5	3	4.20	186	184.80	189
6	7	5.56	182	183.44	189
7	10	7.37	179	181.63	189
8	8	9.49	181	179.51	189
9	9	12.19	180	176.81	189
10	19	17.62	168	169.38	187
Value of test statistic: 6.923 (p = .545)					

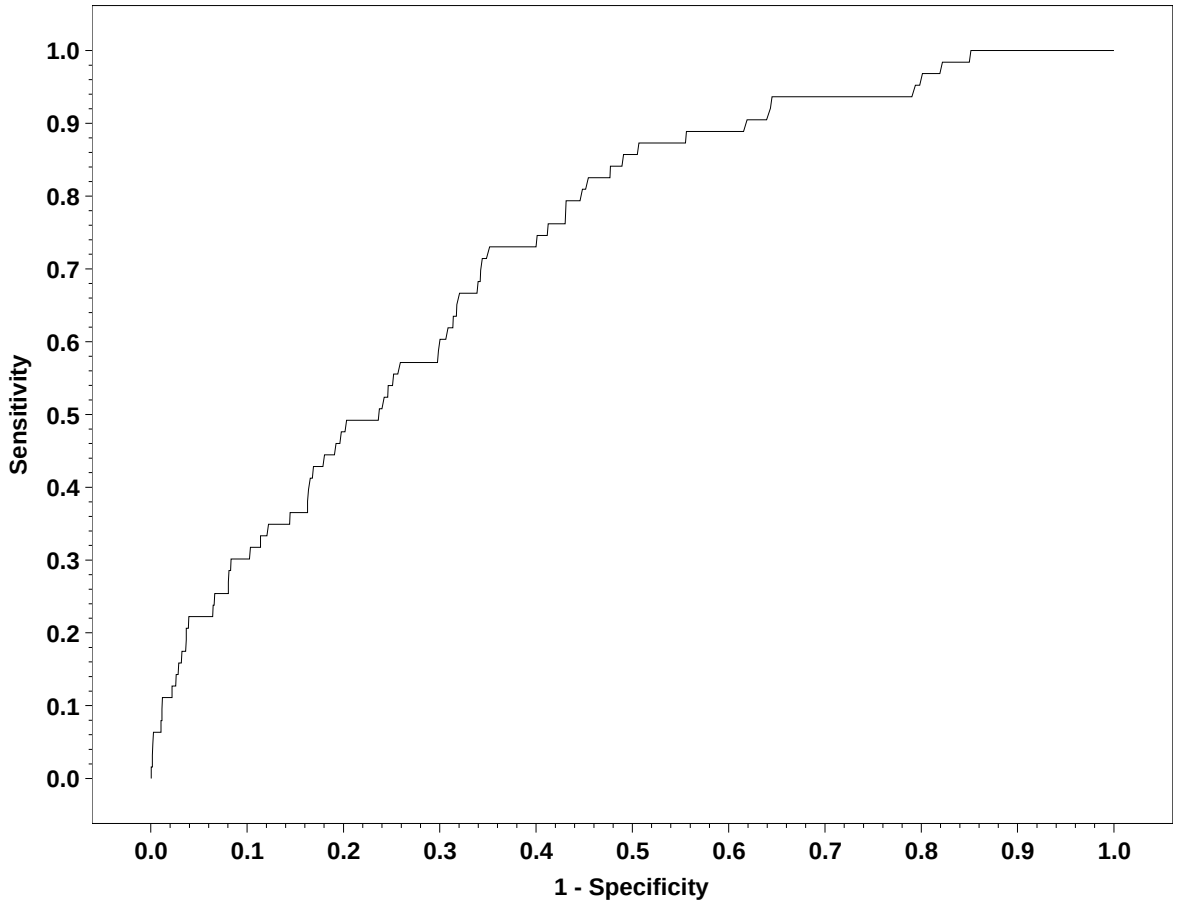
For comparing the predictive ability of logistic regression models, one useful measure is the area under the receiver operating characteristic (ROC) curve, abbreviated AUC. The ROC curve is often used to describe the accuracy of tests in diagnostic medicine, as summarized in the review by Pepe (2000). Briefly, the test yields a numerical result X, such that larger values are more indicative of disease. One can choose a threshold z and dichotomize the test by defining $X \geq z$ as a positive result. From subjects whose true disease status is known (both diseased and nondiseased), one obtains the false-positive rate and the false-negative rate for each value of z. The ROC curve is obtained by plotting 1 minus the false-negative rate against the false-positive rate for all possible choices of z. That is, each value of z yields a point on the curve, which includes the point (0,0) (if z is high enough, the test produces no positives) and the point (1,1) (if

z is low enough, all outcomes are positive). The terms *sensitivity* and *specificity* are also used. The sensitivity, equal to 1 minus the false-negative rate, is the proportion of diseased subjects who are correctly classified. The specificity, equal to 1 minus the false negative rate, is the proportion of nondiseased subjects who are correctly classified. Thus the ROC curve plots sensitivity against 1 minus specificity.

The area under the ROC curve provides a summary of the accuracy of the diagnostic test. As Pepe points out, the AUC "can be interpreted as the probability that the test result from a randomly chosen diseased individual is more indicative of disease than that from a randomly chosen nondiseased individual." This interpretation or equivalence focuses attention on the distributions of the test result (for example, the concentration of a chemical in blood) in diseased and nondiseased persons. If the two distributions are clearly separated, the probability will be close to 1; but if they are centered at the same value, the probability will be ½. In the context of logistic regression one often refers to event cases and non-event cases, rather than diseased and nondiseased persons. The "test result" is the predicted probability of an event, from the logistic regression model.

In this study the "disease" is systemic discrimination. Thus, the area under the ROC curve is the probability that the predicted probability of *SD* (from the model) associated with a randomly chosen establishment with *SD = Y* is greater than that associated with a randomly chosen establishment with *SD = N*. Figure 5.5 shows the ROC curve for the model in Table 5.5. Its AUC is 0.734. This AUC is better than the 0.5 one would get by flipping a coin (corresponding to an ROC curve that runs in a straight line from (0,0) to (1,1)) and within the range (AUC between 0.7 and 0.8) that Hosmer and Lemeshow (2000, Section 5.2.4) characterize as "acceptable discrimination" (between event cases and non-event cases).

Figure 5.5 ROC Curve for the Model of Table 5.5 (predicting the presence of systemic discrimination)



It is also instructive to see how sensitivity and specificity are related to the predicted probability. Figure 5.6 plots these two quantities. This plot facilitates examination of the tradeoff between sensitivity and specificity, if one is trying to choose a cutpoint for predicted probability, above which establishments would be considered likely to be involved in systemic discrimination. Sometimes people choose the value of the predicted probability at which the two curves cross, approximately .04 in Figure 5.6. The plot, however, does not take into account the relative numbers of establishments with *SD = Y* and *SD = N*. Table 5.7 shows the cross-classification that would result (in the data on which the model was based) from using .04 as the cutpoint. The sensitivity and specificity are essentially equal, at 0.67; but the 595 false positives vastly outnumber the 42 true positives. Of the 637 establishments that would be classified as (suspected of having) *SD = Y*, 93% would be false positives.

Figure 5.6 Sensitivity and Specificity versus Predicted Probability for the Model of Table 5.5

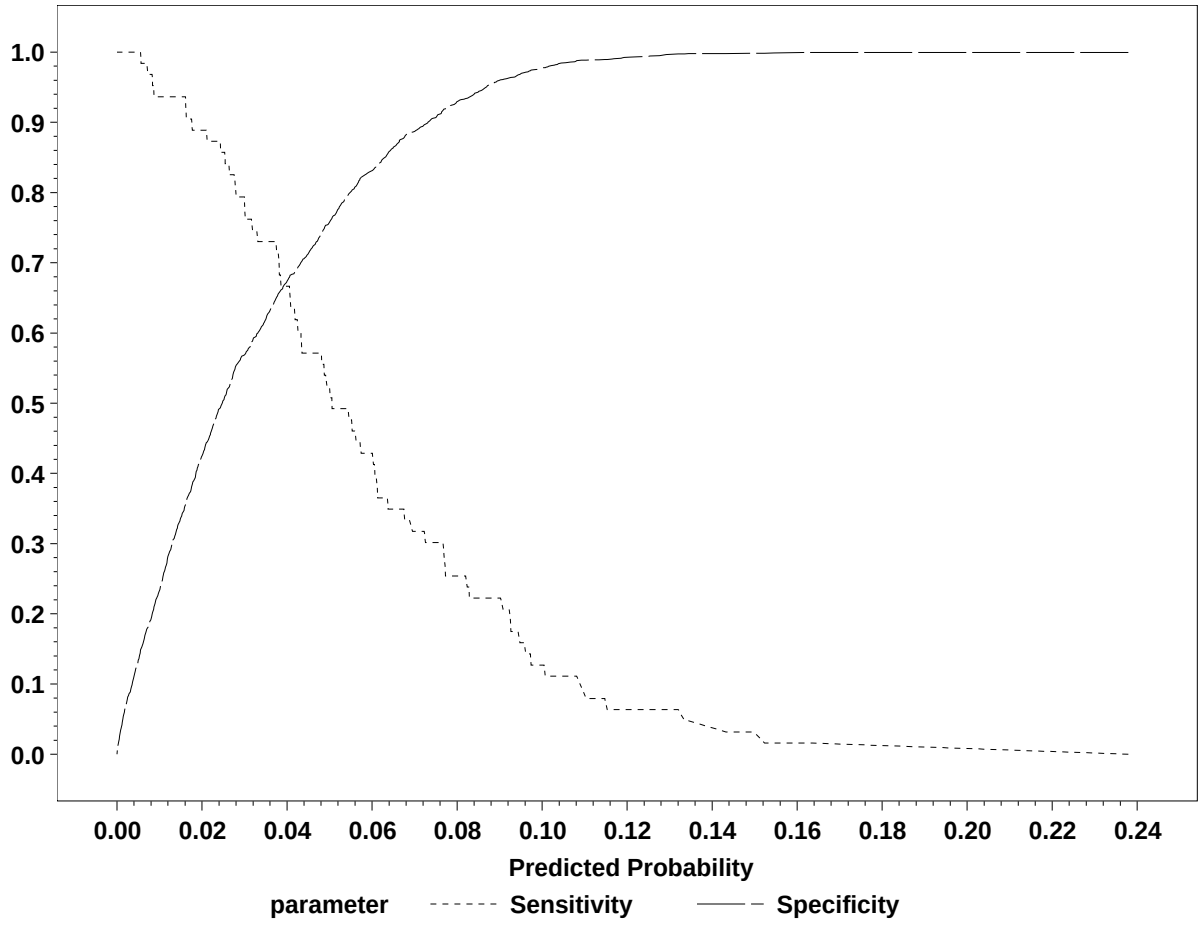


Table 5.7 Classification Table Based on the Model in Table 5.5 Using .04 as the Cutpoint

	Observed		
Classified	***SD = Y***	***SD = N***	**Total**
SD = Y	42	595	637
SD = N	21	1,230	1,251
Total	63	1,825	1,888
Sensitivity = 42/63 = 0.667			
Specificity = 1,230/1,825 = 0.674			

Table 5.8 illustrates the use of a higher cutpoint, .08. Now the specificity would be 0.930, but the sensitivity would be only 0.254. On the other hand, of the 143 establishments that would be classified as *SD = Y*, 16 or 11% would be true positives (somewhat better than the 7% in Table 5.7). It is not immediately clear how to choose an optimal cutpoint. OFCCP could, however use

the predicted probabilities (or, equivalently, the predicted logits, from Table 5.5) to rank establishments for compliance reviews. One approach could give first priority to the establishments with the highest predicted probabilities and continue down the ranking as far as resources permit. An alternative approach would stratify the establishments and oversample the strata with the higher probabilities.

Table 5.8 Classification Table Based on the Model in Table 5.5 Using .08 as the Cutpoint

	Observed		
Classified	***SD = Y***	***SD = N***	**Total**
SD = Y	16	127	143
SD = N	47	1,698	1,745
Total	63	1,825	1,888
Sensitivity = 16/63 = 0.254			
Specificity = 1,698/1,825 = 0.930			

Assessment of the Model by Cross-Validation

When models are fitted to a set of data and measures of the models' predictive ability are calculated from the same set of data, those measures generally rate the models higher than if they were based on fresh data from the same source. If data are abundant, a variety of strategies are available to cope with this inherent shortcoming of the model-building process. One approach, discussed by Hastie et al. (2001, Section 7.2), allocates 50% of the data for fitting the models, uses 25% for selecting among them, and saves the remaining 25% for the final step of assessing the prediction error of the chosen model. The data from completed reviews in the 2002 EO Survey are, however, far from abundant, even after including the establishments whose data were slightly less than "OK." Still, it was possible to carry out some additional assessment of the final model by using a form of cross-validation.

The basic strategy divides the data into ten parts and refits the final model to the ten subsets of the data obtained by leaving out each part in turn. The predicted probabilities for the establishments in each part are then calculated from the model that was fitted with that part omitted. Various summaries based on those predicted probabilities can give an indication of how the model would perform on fresh data.

The 2,153 establishments in the augmented set of data (described in Appendix E) consist of 78 with *SD = Y* and 2,075 with *SD = N*. In order to maintain essentially the same proportion of *SD = Y* in the ten parts, the 78 observations with *SD = Y* were randomly divided into eight groups of 8 and two groups of 7, and the 2,075 observations with *SD = N* were randomly divided into five groups of 207 and five groups of 208. Then each *SD = Y* group was combined with an *SD = N* group, producing seven parts of 215 (five of 8 + 207 and two of 7 + 208) and three parts of 216 (= 8 + 208).

The coefficients of the resulting ten leave-out-one models are shown in Table 5.9. Each of the five coefficients varies somewhat among the ten models, but the variation is not large. The median value of four of the coefficients over the ten models is quite close to the corresponding value from the model fitted to all ten parts ("all," from Table C, in Appendix E). The median for FemMale_Diffi3 is less close, but that coefficient is the least stable in the model. Interestingly, the range of values for most of the coefficients is roughly similar to the coefficient's standard error in the "all" model (Table C); the range for Indicator_GT200 is about half again as large (0.434 versus 0.301). Thus, the ten leave-out-one models do not show dramatic differences.

Table 5.9 Coefficients Obtained by Fitting the Final Logistic Regression Model to the Augmented Data with Each of the Ten Parts Omitted

Part Omitted	Intercept	Indicator_ GT200	MinWhite_ Tenure Ratio	FemMale_ Diffi3	CompFemMale_ TenureRatio
1	–4.809	1.433	–1.519	–8.849	2.335
2	–4.871	0.999	–1.191	–11.113	2.558
3	–4.436	1.508	–1.202	–8.425	1.977
4	–5.198	1.250	–1.538	–10.855	3.002
5	–4.635	1.157	–1.240	–8.923	2.131
6	–4.780	1.134	–1.338	–9.472	2.439
7	–5.008	1.185	–1.099	–10.734	2.451
8	–5.255	1.204	–1.013	–8.681	2.582
9	–5.271	1.036	–1.234	–11.375	3.029
10	–4.893	1.108	–1.098	–8.294	2.344
Median	–4.882	1.146	–1.218	–9.197	2.445
Range	0.835	0.434	0.525	3.081	1.051
All	–4.908	1.150	–1.242	–9.601	2.479

For logistic regression models, one measure of lack of fit (or prediction error) uses, for each observation, –2 times the logarithm of the predicted probability corresponding to the observed outcome in that observation; that is, $-2 \log \hat{p}_i$ if the observed outcome is 1 and $-2 \log (1-\hat{p}_i)$ if the observed outcome is 0. Summing over the observations yields –2 times the log-likelihood of the fitted model. If one divides that sum by the number of observations, the result is the average error of the fitted model. For the model in Table C the value of –2 log *L* is 621.947, and the average error is 0.289. Similarly, as a result of the leave-out-one process, each observation has a predicted probability from the model that omits its part. A calculation analogous to –2 log *L* yields an average value that can be interpreted as an expected prediction error (though it is obtained by cross-validation from the data at hand, rather than from fresh data). In this instance the estimate is 0.293.

One can also use the predicted probabilities from the leave-out-one models to construct an ROC curve and calculate the corresponding AUC. The result is 0.706, somewhat smaller than the 0.722 from the model of Table C but still in the acceptable range.

A Model That Uses Only EEO-1 Data

Among the four predictor variables in the final model (Tables 5.4 and 5.5), two could be constructed from data collected on the EEO-1 form: Indicator_GT200 and FemMale_Diffi3. Fitting that two-variable model gives one indication of the predictive ability that would be possible if only the EEO-1 data were available. Table 5.10 shows the resulting coefficients, standard errors, and p-values for the augmented data (see Appendix E; 2,525 of the 2,534 observations were used).

Table 5.10 Model Based on Two Predictor Variables That Could Be Derived from EEO-1 Data (2,525 observations were used, of which 81 had *SD = Y*)

Variable	Coefficient	Standard Error	P-value
Intercept	−3.891	0.255	<.0001
Indicator_GT200	1.238	0.276	<.0001
FemMale_Diffi3	−9.954	3.170	.0017

The coefficients for the two predictor variables are reasonably close to their values in the four-variable model for the augmented data (in Table C, Appendix E). The AUC for this model is 0.708, slightly lower than the 0.722 for the four-variable model in Table C, but still above the threshold for "acceptable discrimination."

Chapter 6. Summary and Conclusions

The 2002 EO Survey combined a stratified random sample (to ensure coverage of the population of supply and service contractors, after some exclusions) with a systemic random subsample for compliance reviews, focusing on systemic discrimination. This design avoided one of the major problems of the previous studies and provided a solid statistical basis for investigating the relation between systemic discrimination and characteristics of contractor establishments, obtained from the EO Survey data.

Of the 10,018 establishments in the sample for the 2002 EO Survey, 85% responded in some way. A substantial number (27% of the sample), however, contested the OFCCP's jurisdiction, and others (15%) submitted data that were not satisfactory, leaving 4,254 surveys whose data were "OK". However, even within those surveys the quality of the data varied widely; some items produced outliers or inconsistencies that went undetected when the data were transferred from the survey forms to the database. Through statistical analysis it was possible to control the impact of such data problems for the purposes of this study. Within the subsample of 6,400 establishments selected in advance for compliance reviews, 3,048 had surveys that were "OK," including 2,226 with completed reviews. Among those with completed reviews 67 (3.0%) had findings of systemic discrimination.

The data from the 4,254 "OK" surveys were used to develop a total of 125 predictor variables, drawing on all items in Part B and Part C of the EO Survey. Several steps of analysis and model building, starting with the full set of predictors, produced a logistic regression model that related the presence or absence of systemic discrimination to four predictor variables:

- Indicator_GT200, whether the establishment reported more than 200 full-time employees;

- MinWhite_TenureRatio, the average (over EEO-1 categories) of the ratio of average tenure among minority employees to average tenure among non-minority employees;

- FemMale_Diffi3, the absolute value of the difference between the proportion of female employees and the proportion of male employees in EEO-1 Category 3 (technicians);

- CompFemMale_TenureRatio, the average (over EEO-1 categories) of the ratio of the female-to-male tenure ratio to the median of those ratios in the establishment's comparison group.

Although this model fit the data reasonably well and had acceptable predictive ability (as indicated by the area under its ROC curve, 0.734), models tend to be "tuned" to the data that are used in fitting them, and so measures of their performance may be optimistic. Assessment of the model by cross-validation, however, indicated that the tuning effect was not serious.

The low prevalence of systemic discrimination in the population of supply and service contractors, and its relation to some of the predictor variables, however, limit the usefulness of the model and the survey.

- Systemic discrimination was found in only about 3% of the establishments reviewed, and those with findings of *SD* did not share any combinations of characteristics that set them apart from establishments with findings of no *SD*. Thus, screening on the basis of the predicted probabilities would be expected to produce large numbers of false positives.

- The directions of the contributions of FemMale_Diffi3 and CompFemMale_TenureRatio are counterintuitive, and those same directions are present in the separate relations of *SD* to those variables: establishments with findings of *SD* tended to have smaller values of FemMale_Diffi3 and larger values of CompFemMale_TenureRatio than establishments with findings of no *SD*. These results contrast with the seemingly intuitive directions of the other two predictors: findings of *SD* were more prevalent among establishments with more than 200 employees, and establishments with findings of *SD* tended to have smaller values of MinWhite_TenureRatio.

The ability to use a model and to use data from the EO Survey may be strengthened by more extensive editing and cleaning of submitted data, before they are incorporated in any EO Survey database. The cleaning procedures should focus on responses that are likely to be invalid and on inconsistencies in the responses.

An Alternative Approach

OFCCP and Abt Associates discussed the feasibility of an alternative approach for additional analysis and development of a targeting model. For example, OFCCP could select a stratified random sample of establishments for compliance reviews. During the compliance reviews, OFCCP personnel could use the data provided by contractors at the desk audit stage to develop specified data elements. The compliance reviews would proceed under normal OFCCP protocols. Over several years OFCCP could accumulate a substantial amount of data, consisting of the results of compliance reviews for particular establishments and corresponding data elements similar to those collected by the EO Survey. This approach has the advantage of collecting more-accurate and more-pertinent data than provided by the current EO Survey, and OFCCP could use the database for additional study through the techniques described in this report.

A related advantage is that OFCCP would avoid the expense of the survey process, including costs involved with "cleaning" data. It also seems likely that cleaned data would be available for analysis sooner than is possible with the EO Survey.

Another advantage is that OFCCP can ensure that the data collected for an establishment come from a time period over which OFCCP will assess the employer's personnel practices. For example, OFCCP will be able to ensure that it collects data on applicants and hires for the same period over which it reviews the contractor's hiring practices for potential discrimination.

As in the present study, the randomness of the sample would be important, as the basis for inferences from the sample of establishments to the universe of contractors. Some degree of stratification (e.g., on size of establishment) would probably be worthwhile.

References

Bendick, Marc Jr. and Miller, John J. (2000), Equal Opportunity Survey: Documentation of a Preliminary Analysis Delivered to OFCCP on June 21, 2000. Washington, DC: Bendick and Egan Economic Consultants, Inc., June 21, 2000.

Bendick, Marc Jr., Miller, John J., Blumrosen, Alfred J., and Blumrosen, Ruth Gerber (2000), The Equal Opportunity Survey: Analysis of a First Wave of Survey Responses. Washington, DC: Bendick and Egan Economic Consultants, Inc., September 2000.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2001), *The Elements of Statistical Learning*. New York: Springer-Verlag.

Hosmer, David W. and Lemeshow, Stanley (2000), *Applied Logistic Regression*, second edition. New York: John Wiley & Sons.

Office of Federal Contract Compliance Programs (2000), Report on the Use of the Equal Opportunity Survey, Prepared for the Office of Management and Budget (OMB). Washington, DC: Office of Federal Contract Compliance Programs, Employment Standards Administration, U.S. Department of Labor, October 2000.

Pepe, Margaret Sullivan (2000), Receiver Operating Characteristic Methodology. *Journal of the American Statistical Association* 95, 308-311

Appendix A

Frequency Distributions
for the Sampling Frame, Sample, and Subsample

Appendix B

EO Survey Instrument

Appendix C

Comparison Groups for the Comparative Variables

Appendix D

List of Predictor Variables

Appendix E

Memorandum on Issues Arising from OFCCP's
Review of the Draft Report